Transformer les collections en information grâce aux technologies du web sémantique

Etienne Cavalié, Géraldine Geoffroy

Transformer les collections en information grâce aux technologies du web sémantique

Etienne Cavalié, Géraldine Geoffroy

Arabesques (1269-0589), n° 80, oct.-nov.-déc. 2015, p. 18-19

Nos catalogues sont conçus, prévus, pour nous permettre de décrire nos collections, de les gérer et d'y donner accès. Cette origine a des conséquences sur les choix techniques et normatifs que nous faisons. Par exemple, nous mutualisons autant que possible la description bibliographique, identique ; mais nous n'avons pas de raison (sauf pour le PEB) de partager nos données d'exemplaires et encore moins nos données de gestion (liées aux étapes de la commande, par exemple). Les bases de données se sont donc juxtaposées côte à côte, parce qu'aucun usage prévu ne justifiait une autre approche. Hors de la recherche et de la gestion de documents, nos données en tant que telles ne servent (quasiment) à rien d'autre.

Nos données ont une valeur d'usage : c'est-à-dire qu'elles ont d'autant plus de valeur qu'elles « servent », qu'elles sont un apport de connaissance ou une aide à la décision. Si nous améliorons les conditions de réutilisation, si nous facilitons les possibilités de réutilisation, leur valeur grandit et nous nous inscrivons plus solidement dans le *LOD Cloud diagram*.

Car le nouvel environnement du web, et les technologies du web de données et du *linked data*, nous invitent à porter sur elles un autre regard : les concevoir non plus comme des *méta*données (c'est-à-dire des données invitant aussitôt à détourner d'elles le regard, pour le porter sur l'objet qu'elles décrivent), mais comme des données, des objets exploitables en tant que tels, ayant un intérêt et une valeur intrinsèques.

Nos bases sont certes la description de nos collections, mais elles sont aussi une masse d'informations considérables. Informations sur quoi ? Sur la collection et sur sa description : mais ces deux notions s'élargissent rapidement :

- à la bibliothèque qui a constitué ces collections (histoire et disciplines de l'établissement, idéologie sous-jacente des bibliothécaires qui ont choisi ou non d'acheter précocement des livres sur le développement durable ou les *gender studies*, etc.),
- à la manière de les classer et les indexer
- aux auteurs qui ont écrit ces ouvrages,
- à la société qui a produit ces objets éditoriaux,
- aux usages qui en sont faits (statistiques de prêts notamment, mais pas seulement)

Bref, il est possible de concevoir la collection comme un artéfact produit par un ensemble d'acteurs, et à ce titre fournissant des informations sur les acteurs des différentes étapes.

Par exemple, pour étudier l'évolution des intérêts de la recherche scientifique française (et à travers elle, ou comparée à elle, de la société française entière) concernant le Moyen-Orient, les thèses universitaires et les mots présents dans les titres, les résumés, les sujets, sont une source riche. L'étude du langage d'indexation RAMEAU, et son évolution au fil des décennies, sont un reflet des modalités de catégorisation de thématiques contemporaines : le féminisme, la laïcité, etc... qui en outre évoluent au cours du temps.

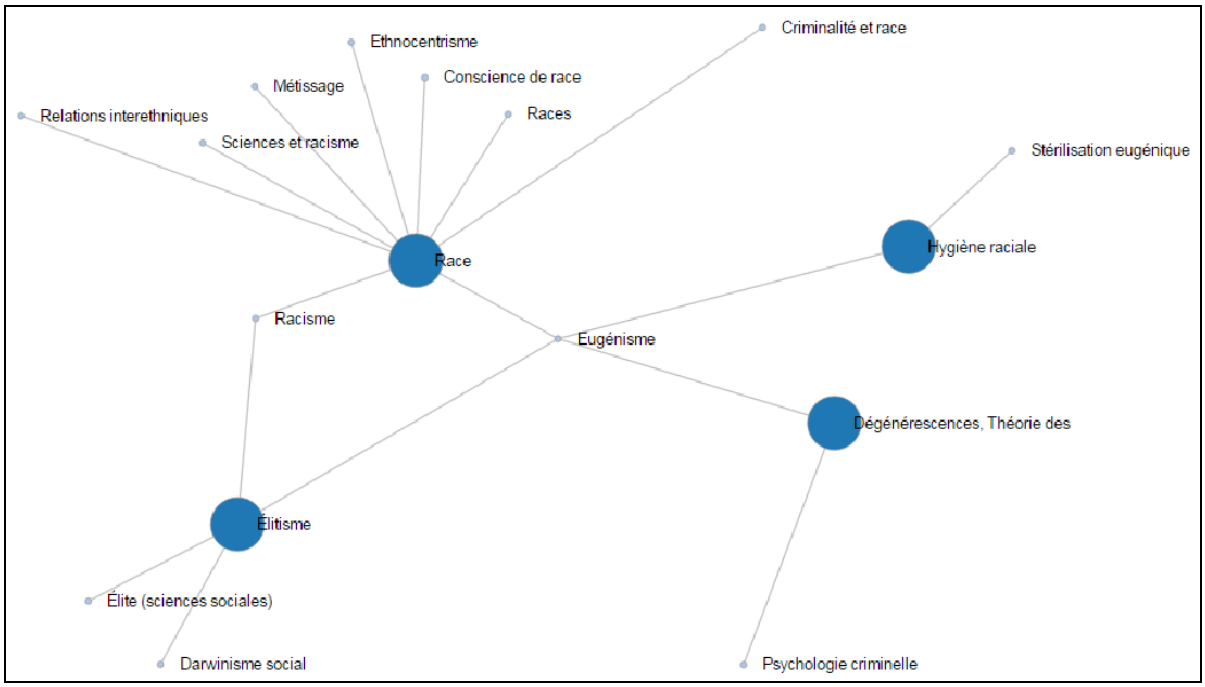

Figure 1 : l'eugénisme selon RAMEAU

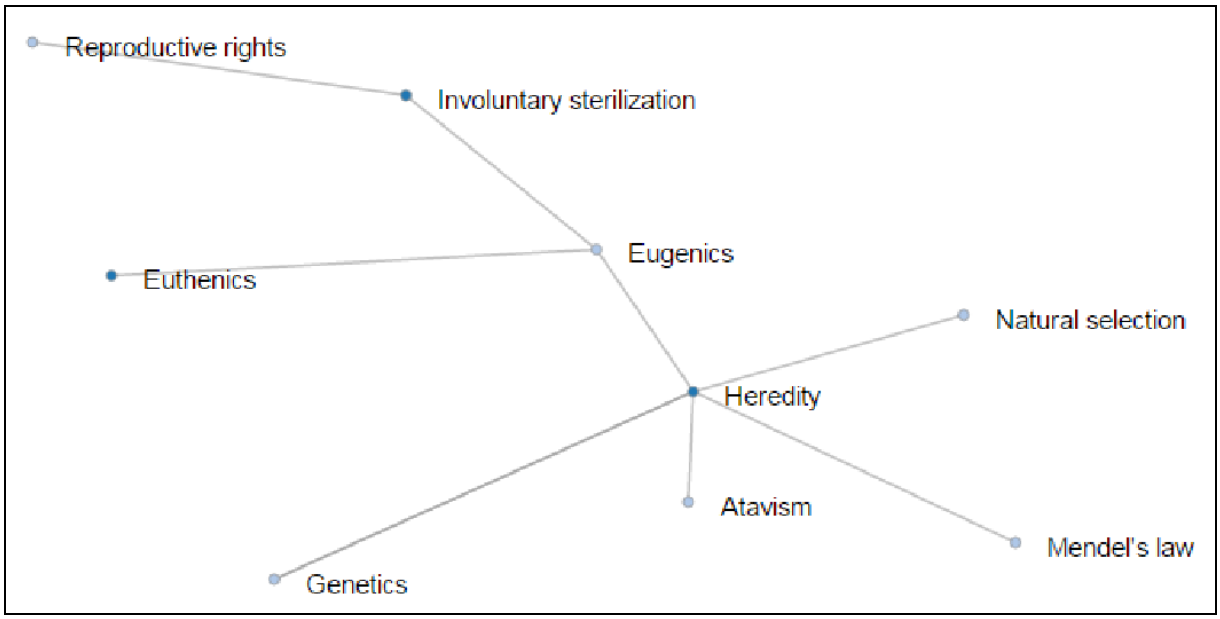

Figure 2 : l'eugénisme selon la LCC

La description de la collection renvoie donc à un ensemble d'objets, individus, structures, qui existent « dans le monde réel ». Mais dans un premier temps, on constate que les données présentes dans cette description sont refermés sur eux-mêmes : les PPN d'autorités ne servent qu'à décrire les notices du Sudoc (ou à peine plus), et réciproquement.

L'enjeu majeur est donc que nos données renvoient aussi aux manières de désigner plus « normalement » les objets qu'elles décrivent. Et c'est tout ce qu'on place derrière l'expression « alignement des référentiels » : La nature des informations que peut alors être amenée à diffuser une bibliothèque n'est plus biblio-centrée lorsqu'elle se lie à des concepts qui lui sont extérieurs, lorsque les données bibliographiques sont liées à des données encyclopédiques (DBPedia, Freebase…), statistiques (INSEE), géographiques (Geonames, IGN)...qui associent des données à des objets, des lieux, des personnes.

La production bibliographique, par exemple, vient enrichir automatiquement l'ensemble des connaissances produite sur une aire géographique (et l'on découvrira alors que les auteurs nés dans telles régions réagissent plutôt positivement ou négativement aux questions d'immigration, par exemple).
Les technologies du *linked data*, les données structurées en RDF, sont précisément conçues pour permettre l'élaboration d'une connaissance qui se construit par associations de concepts, entre bases de données multiples et diverses.
C'est ce processus de liage qui fait qu'une base bibliographique se transforme en ensemble d'informations : en étant lié aux concepts extérieurs à la collection elle-même. Et ces liens ne sont pas seulement ceux de bases d'autorités (celles q servent à indexer, ou à désigner une personne de manière univoque), mais vers les personnes ou les objets eux-mêmes : à savoir VIAF/DBpedia/LinkedIn, etc.
En prenant le temps « d'aligner les référentiels », on facilite le parcours, les rebonds, les mash-ups. Nos données sont exploitables, au même titre que plein d'autres comme masse d'informations sur la société qui les a produites, ou qui en fait usage.
Dans cette démarche, deux aspects ne doivent pas être bloquants :
D'abord, les bibliothèques ne produisent et ne gèrent pas seulement des données bibliographiques. Les données de gestion, les exemplaires, les statistiques de prêts, les catégories de populations de lecteurs, les coûts de la documentation, les vols de livres, les horaires d'ouverture, sont autant d'informations à adjoindre aux données bibliographiques, pour comprendre les conditions d'usage et d'accès des collections décrites.
Ensuite, si on regarde le classement par étoiles de Tim Berners-Lee, la difficulté de RDFiser nos données (« Comment dit-on *Nombre de prêts* en RDF ? ») ne doit pas nous empêcher de viser la cinquième étoile, y compris en faisant abstraction de la quatrième.

Figure 3 : les 5 étoiles du Linked Data (source : http://www.w3.org/DesignIssues/LinkedData.html)

Aligner trois référentiels, par exemple les identifiants ESGBU des Universités avec leurs identifiants PAPESR et leurs identifiants DBPedia, peut pour commencer se faire dans un fichier CSV, et rendre des services pour croiser les informations issues de ces différentes bases.

Une fois que les données sont produites, le choix de l'endroit où les déposer est déjà une forme de recontextualisation, et une incitation à certaines réutilisations : serveur local ou plate-forme régionale, portail d'open data du Ministère de l'Enseignement supérieur ou data.gouv.fr.
Il est inutile de vouloir anticiper ce que d'autres pourront bien faire de ces données. En revanche on peut chercher à être le premier bénéficiaire de cette exposition. Ainsi les bibliothèques de la NCSU publient en SKOS une base de fournisseurs de ressources électroniques pour alimenter leur ERMS (développé localement). Plutôt que d'enfermer dans leur logiciel des fiches décrivant les fournisseurs, éditeurs, etc. ils ont directement alimenté une base ouverte où toute application (à commencer par les leurs) puisse venir puiser à tout moment. Mais les fiches fournisseurs que je gère dans mon SIGB ne sont pas liées à celles de Sifac.
Imaginons quelques utilisations tout de même : les statistiques de prêt des bibliothèques permettent de mesurer la « proximité » entre deux documents, et proposer en open data cette « proximité » serait un service de recommandation comparable à celui de certains sites web « Ceux qui ont acheté… ont aussi acheté… ». Cette base pourrait alimenter une application d'« activités culturelles » qui suggèrerait, pour un samedi après-midi libre, une visite au musée, un cinéma ou une lecture ; ce peut-être aussi des suggestions faites à l'issue d'une commande de billet de train ou d'avion : pour occuper votre voyage, téléchargez tel podcat, lisez tel livre.
Certes, le livre lu sera peut-être téléchargé, ne sera peut-être pas l'exemplaire acquis pas la bibliothèque. Mais il sera recommandé grâce aux données de prêt de la bibliothèque.
La plus grande difficulté réside peut-être là : admettre que nos données, nos précieuses données, soient mêlées à d'autres ; et qu'elles soient utilisées sans que nos collections le soient.
Mais pourquoi ne pas considérer que nos données sont aussi des collections ?